Modeling relationships in referential expressions with compositional modular networks

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017
Hu, R., Rohrbach, M., Andreas, J., Darrell, T., & Saenko, K.

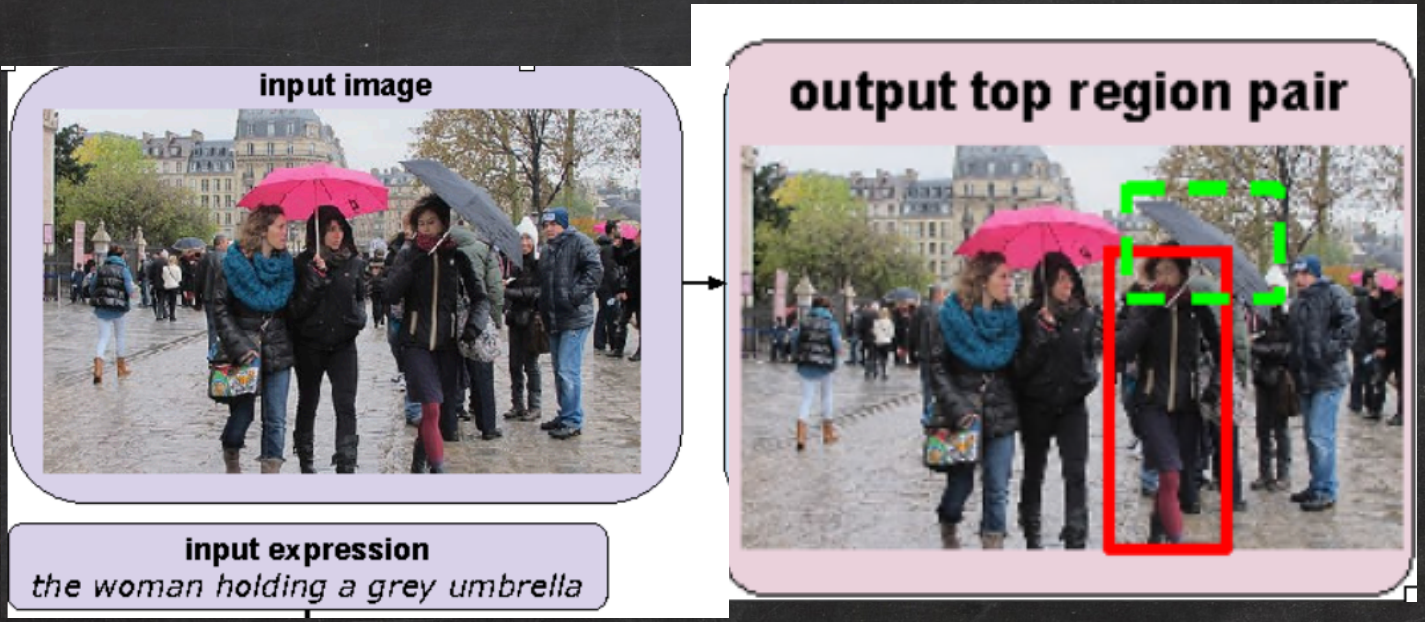
Presenter: Minzhi Liao
Hooman Hashemi

- Introduction
  - Motivation
  - Related work
- Model in the paper
  - Expression parsing
  - Localization module
  - Relationship module
- Experiments
  - Synthetic dataset
  - Visual Genome dataset
  - Google-Ref dataset
  - Visual-7W dataset

Outline

- Introduction
  - Motivation
  - Related work
- Model in the paper
  - Expression parsing
  - Localization module
  - Relationship module
- Experiments
  - Synthetic dataset
  - Visual Genome dataset
  - Google-Ref dataset
  - Visual-7W dataset

Motivation



Problem: given a referential expression and image localize the entities

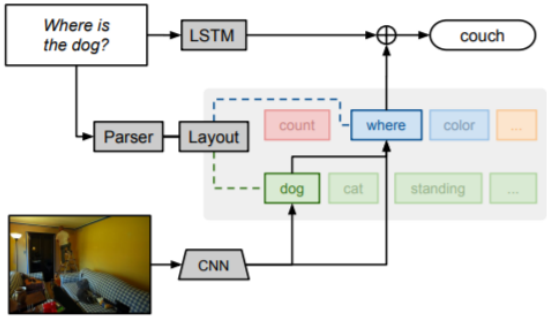
Related work

Deep Compositional Question Answering with Neural Module Networks

Jacob Andreas Marcus Rohrbach Trevor Darrell Dan Klein
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
{jda,rohrbach,trevor,klein}@{cs,eecs,eecs,cs}.berkeley.edu

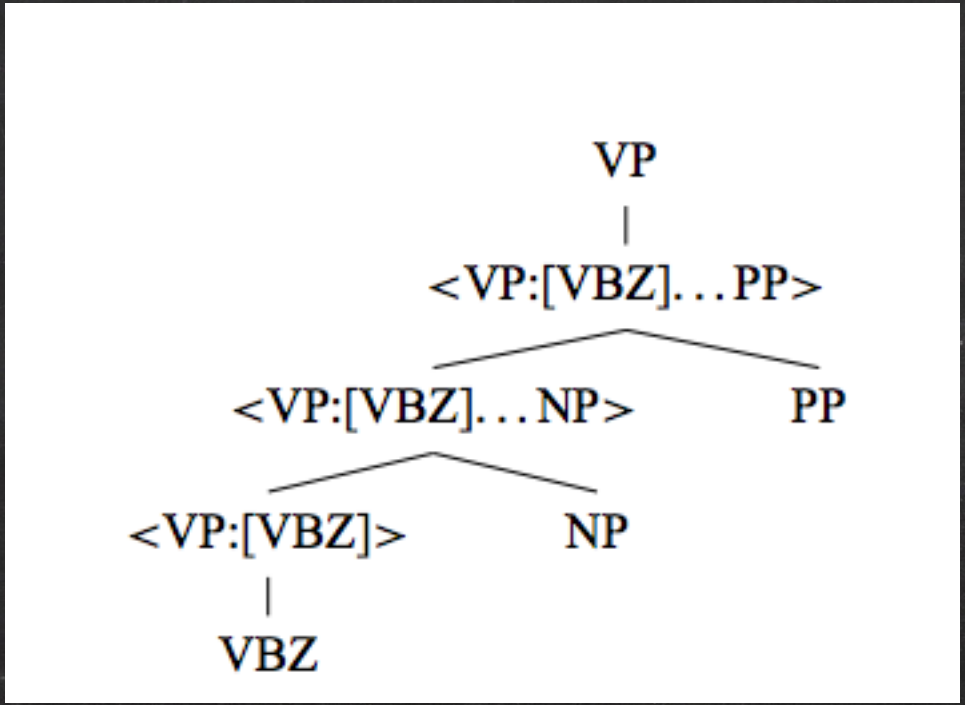
Abstract

Visual question answering is fundamentally compositional in nature—a question like where is the dog? *shares substructure with questions like* what color is the dog? *and* where is the cat? *This paper seeks to simultaneously exploit the representational capacity of deep networks and the compositional linguistic structure of questions. We describe a procedure for constructing and learning* neural module networks, *which compose collections of jointly-trained neural "modules" into deep networks for question answering. Our approach decomposes questions into their linguistic sub-*



Main idea: dynamic assembling network architecture with modules for simple tasks

Layout (Parsing)

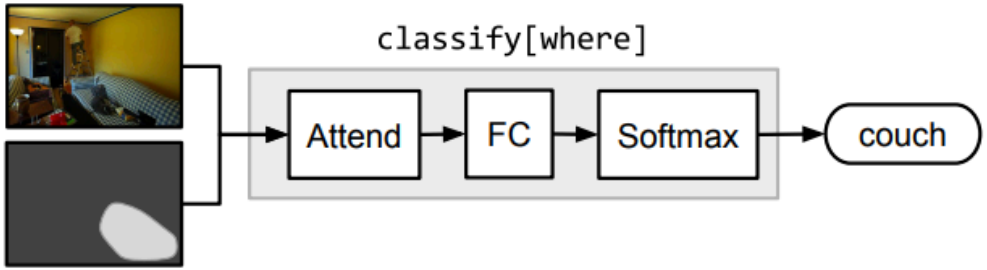


Main idea: dynamic assembling network architecture with modules for simple tasks

Modules

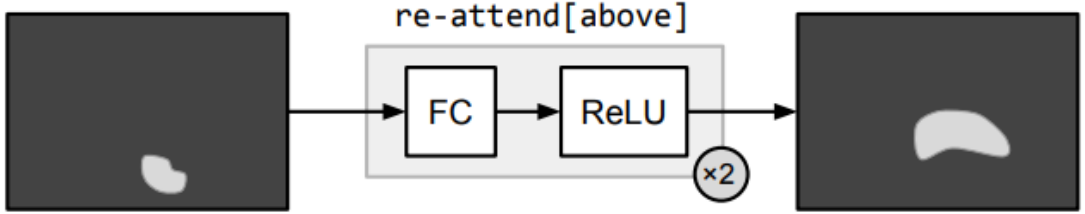
Classification

$$\texttt{classify} : \mathit{Image} \times \mathit{Attention} \rightarrow \mathit{Label}$$

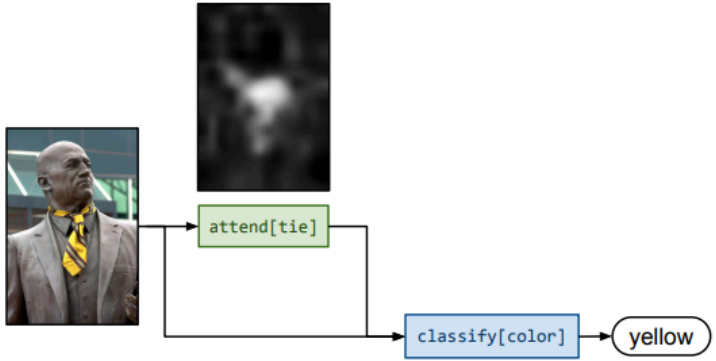


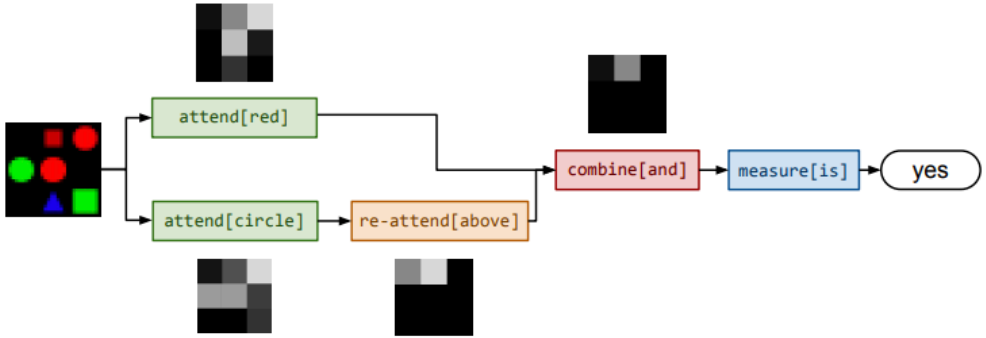
Re-attention

$$\texttt{re-attend} : \mathit{Attention} \rightarrow \mathit{Attention}$$



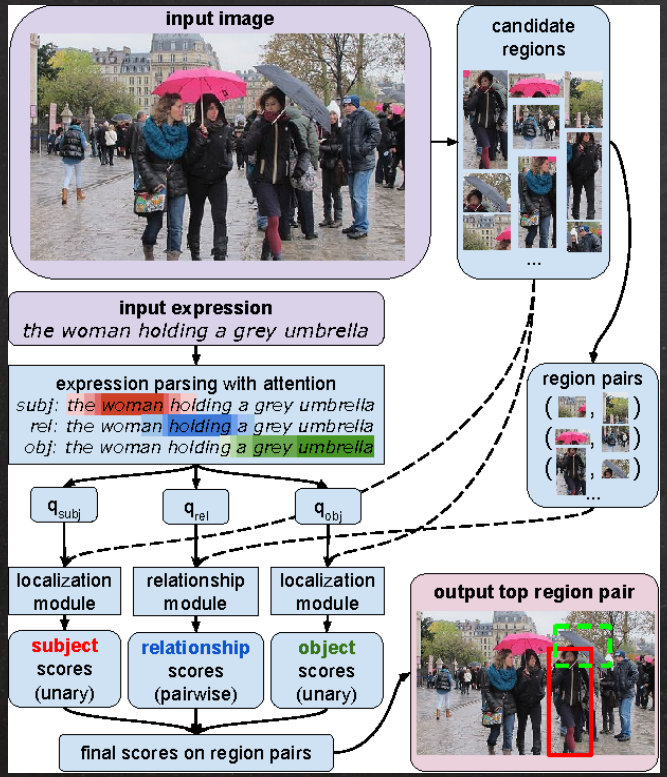
Example





Main idea: dynamic assembling network architecture with modules for simple tasks

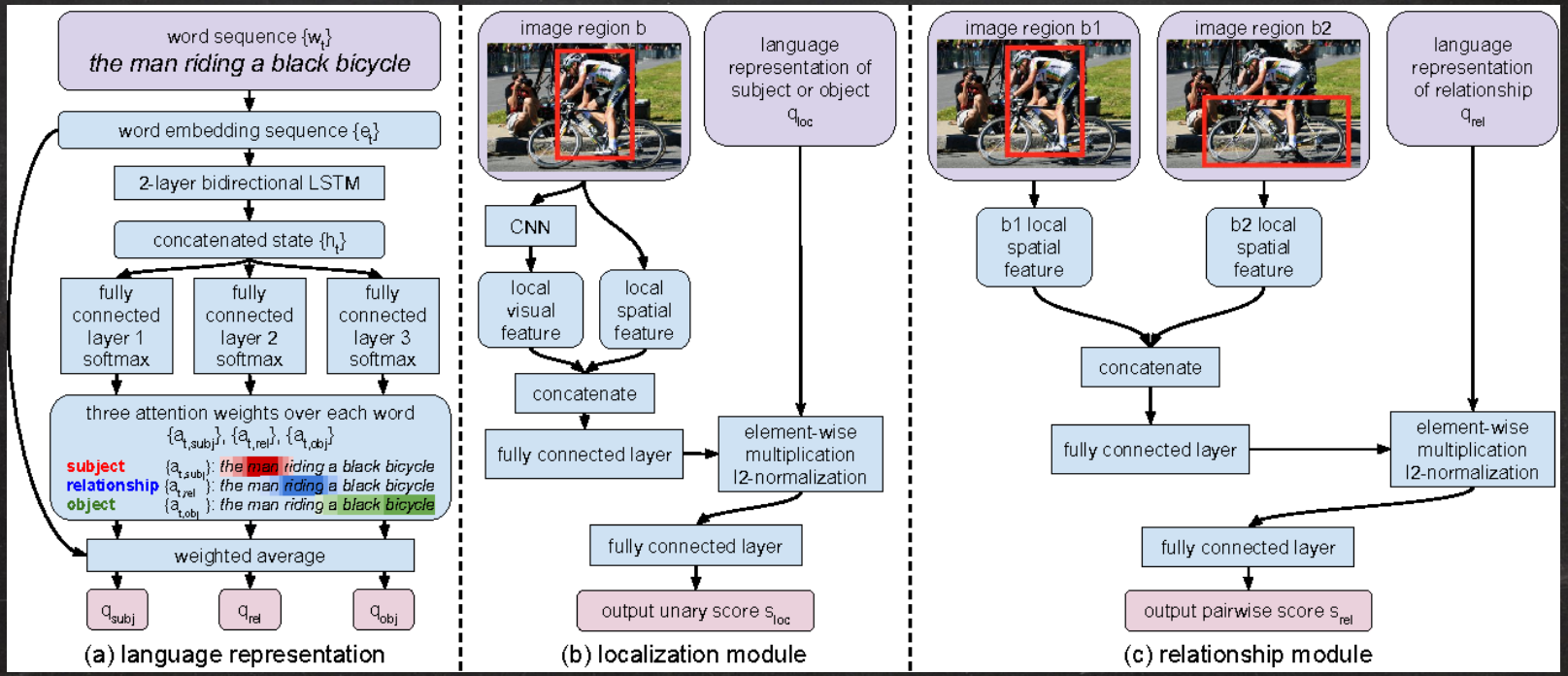
Motivation



Localizing entities based on arbitrary natural language expression is a challenging problem.

Previous work either treat referential expression holistically, or relies on a fixed set of entity and relationship categories.

Model in the paper



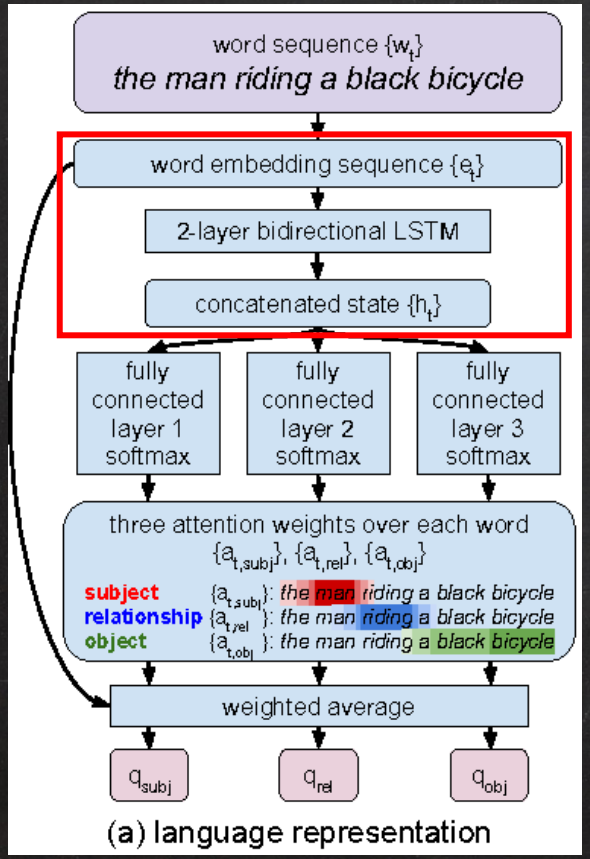
Expression parsing

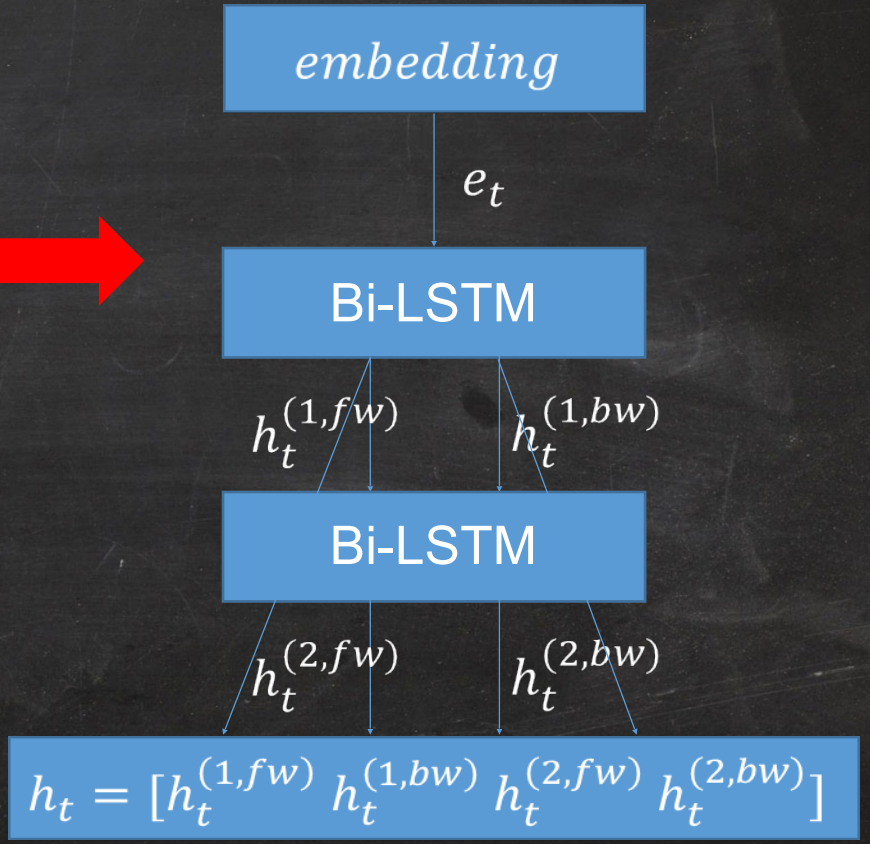
the apple on top of the bookshelf

the apple on top of the bookshelf

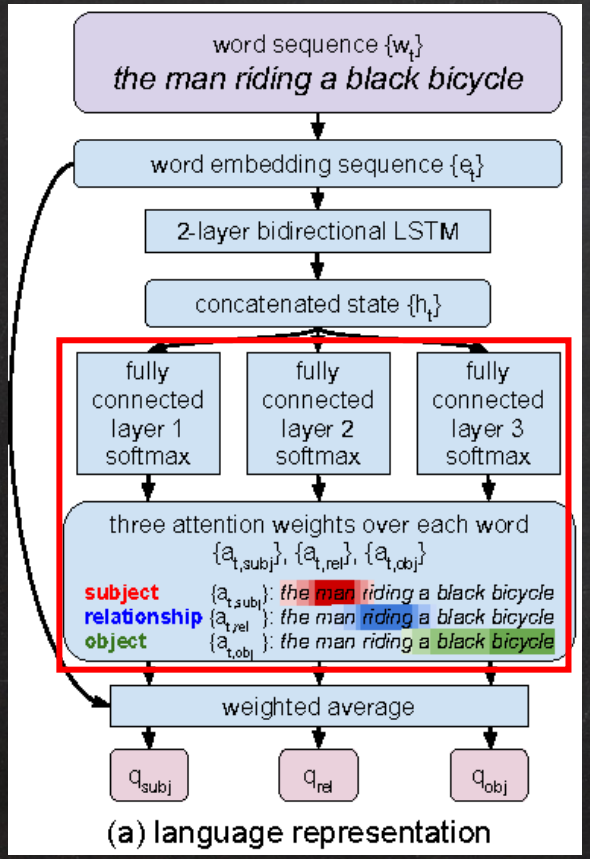
Results from syntactic parsers not always correspond to intuitive visual representations

Expression parsing





Expression parsing



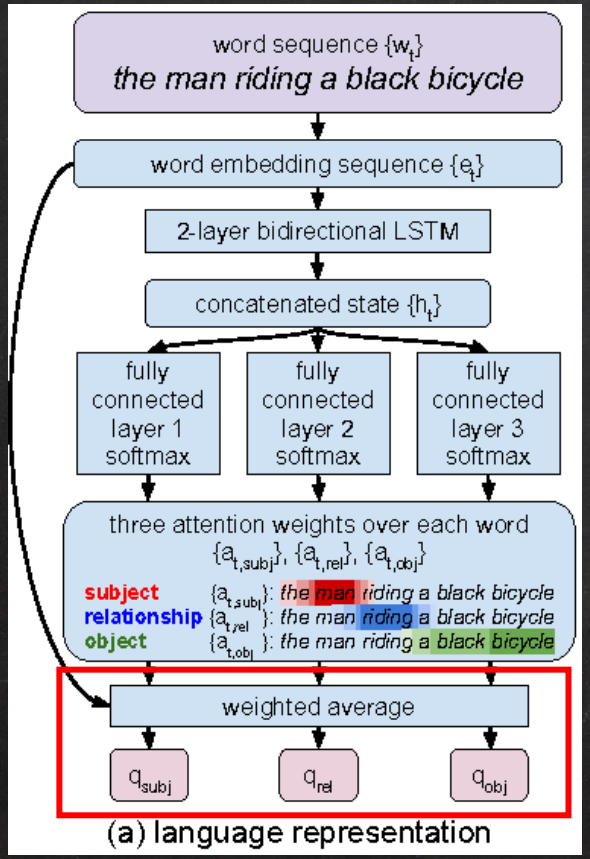
(a) language representation

$$a_{t,subj} = \exp\left(\beta_{subj}^T h_t\right) / \sum_{\tau=1}^{T} \exp\left(\beta_{subj}^T h_\tau\right)$$

$$a_{t,rel} = \exp\left(\beta_{rel}^T h_t\right) / \sum_{\tau=1}^{T} \exp\left(\beta_{rel}^T h_\tau\right)$$

$$a_{t,obj} = \exp\left(\beta_{obj}^T h_t\right) / \sum_{\tau=1}^{T} \exp\left(\beta_{obj}^T h_\tau\right)$$

Expression parsing

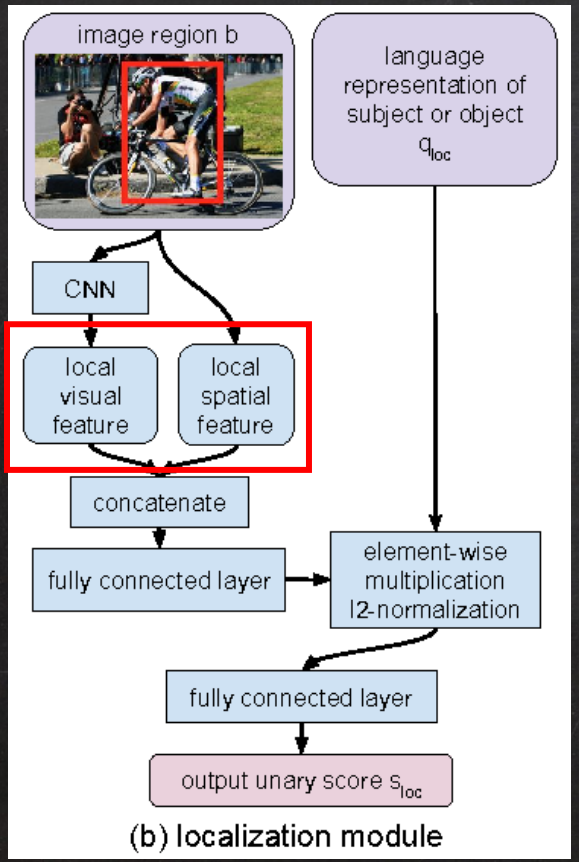


$$q_{subj} = \sum_{t=1}^{T} a_{t,subj} e_t$$

$$q_{rel} = \sum_{t=1}^{T} a_{t,rel} e_t$$

$$q_{obj} = \sum_{t=1}^{T} a_{t,obj} e_t.$$

Localization module



Local spatial feature:

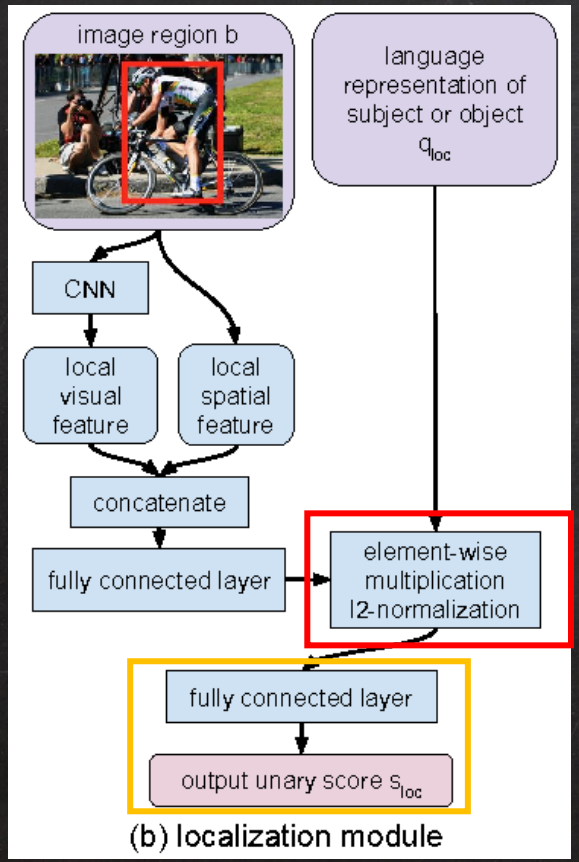
$$x_s = \left[\frac{x_{min}}{W_I}, \frac{y_{min}}{H_I}, \frac{x_{max}}{W_I}, \frac{y_{max}}{H_I}, \frac{S_b}{S_I}\right]$$

$x_{min}, y_{min}, x_{max}, y_{max}$: coordinates of bounding box

W_I, H_I: width and height of image

S_I, S_b: area of image and bounding box

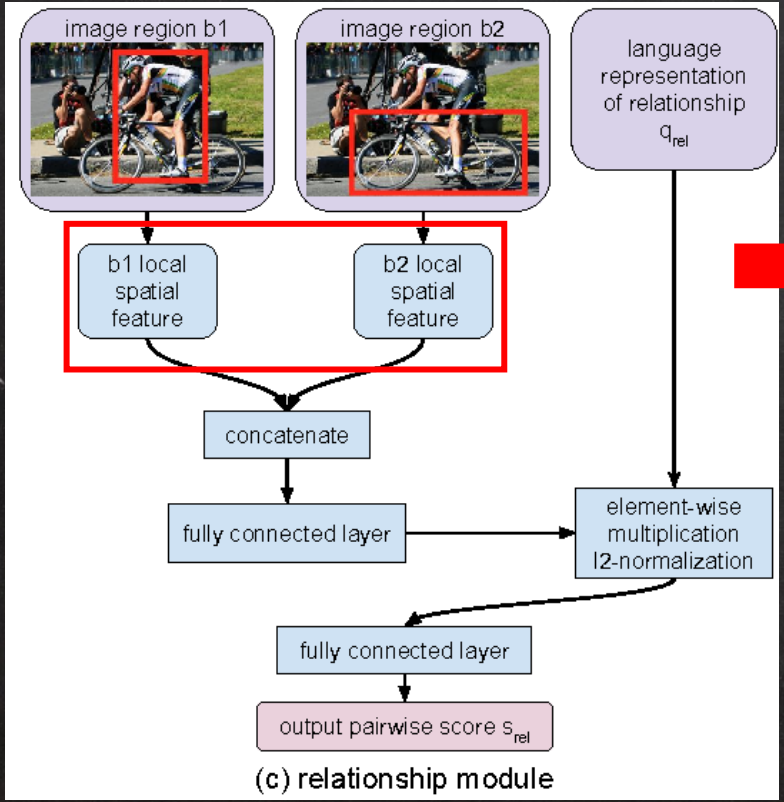
Localization module



$$\begin{aligned} \tilde{x}_{v,s} &= W_{v,s}x_{v,s} + b_{v,s} \\ z_{loc} &= \tilde{x}_{v,s} \odot q_{loc} \\ \hat{z}_{loc} &= z_{loc}/\|z_{loc}\|_2 \end{aligned}$$

$$s_{loc} = w_{loc}^T \hat{z}_{loc} + b_{loc}.$$

Relationship module



(c) relationship module

Local spatial feature:

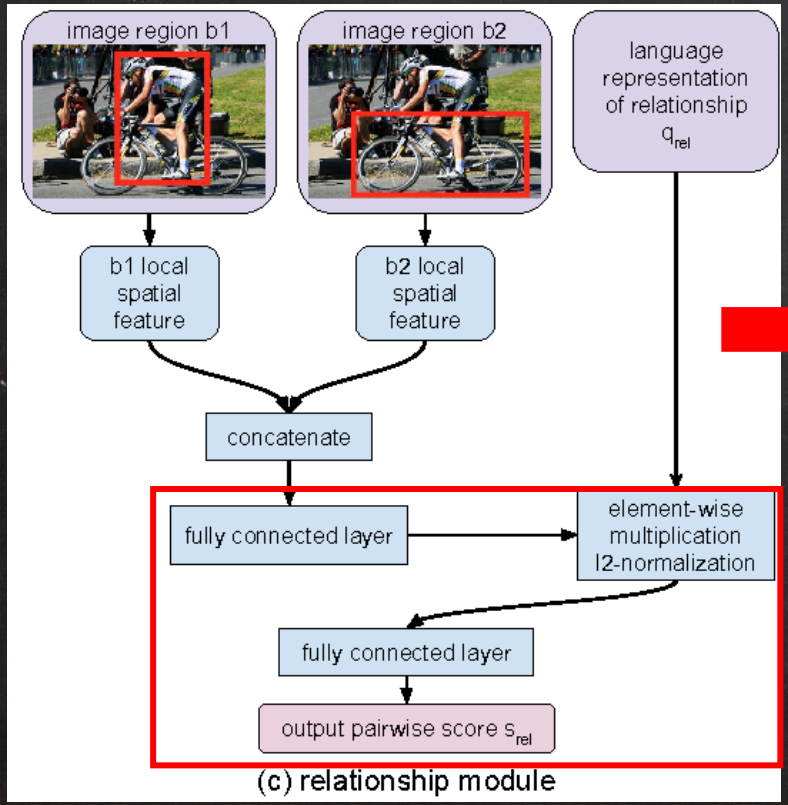
$$x_s = \left[\frac{x_{min}}{W_I}, \frac{y_{min}}{H_I}, \frac{x_{max}}{W_I}, \frac{y_{max}}{H_I}, \frac{S_b}{S_I}\right]$$

$x_{min}, y_{min}, x_{max}, y_{max}$: coordinates of bounding box

W_I, H_I: width and height of image

S_I, S_b: area of image and bounding box

Relationship module



$$
\begin{aligned}
\tilde{x}_{s1,s2} &= W_{s1,s2} x_{s1,s2} + b_{s1,s2} \\
z_{rel} &= \tilde{x}_{s1,s2} \odot q_{rel} \\
\hat{z}_{rel} &= z_{rel} / \|z_{rel}\|_2 \\
s_{rel} &= w_{rel}^T \hat{z}_{rel} + b_{rel}.
\end{aligned}
$$

Loss function

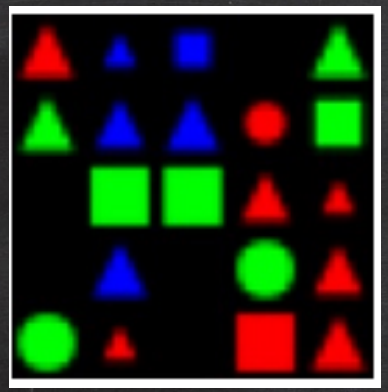
Ground truth subject and object

$$Loss_{strong} = -\log\left(\frac{\exp\left(s_{pair}(b_{subj\_gt}, b_{obj\_gt})\right)}{\sum_{(b_i,b_j)\in B\times B}\exp\left(s_{pair}(b_i, b_j)\right)}\right)$$

Ground truth subject only

$$Loss_{weak} = -\log\left(\frac{\exp\left(s_{subj}(b_{subj\_gt})\right)}{\sum_{b_i\in B}\exp\left(s_{subj}(b_i)\right)}\right)$$

Synthetic dataset



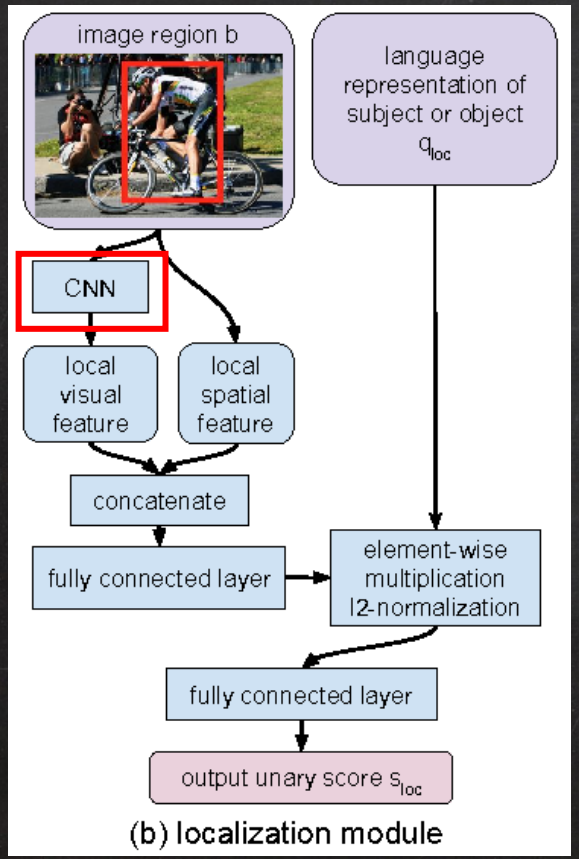
Shapes in different colors and size

5 by 5 grid

expression=*“the green square right of a red circle”*

[subject] [relationship] [object]

Experiments on synthetic dataset



(b) localization module

VGG-16 pretrained on ImageNet classification

Bounding box proposals are 5 by 5 grids

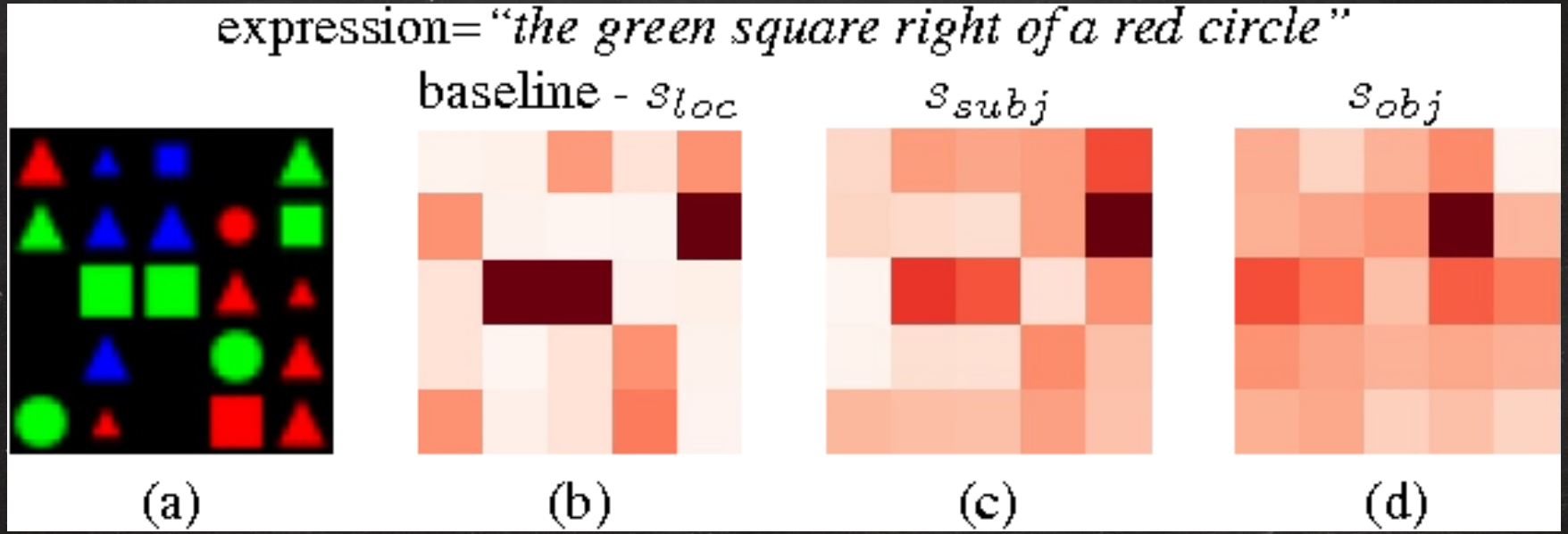
Experiments on synthetic dataset

Method	Accuracy
baseline (loc module)	46.27%
our full model	99.99%

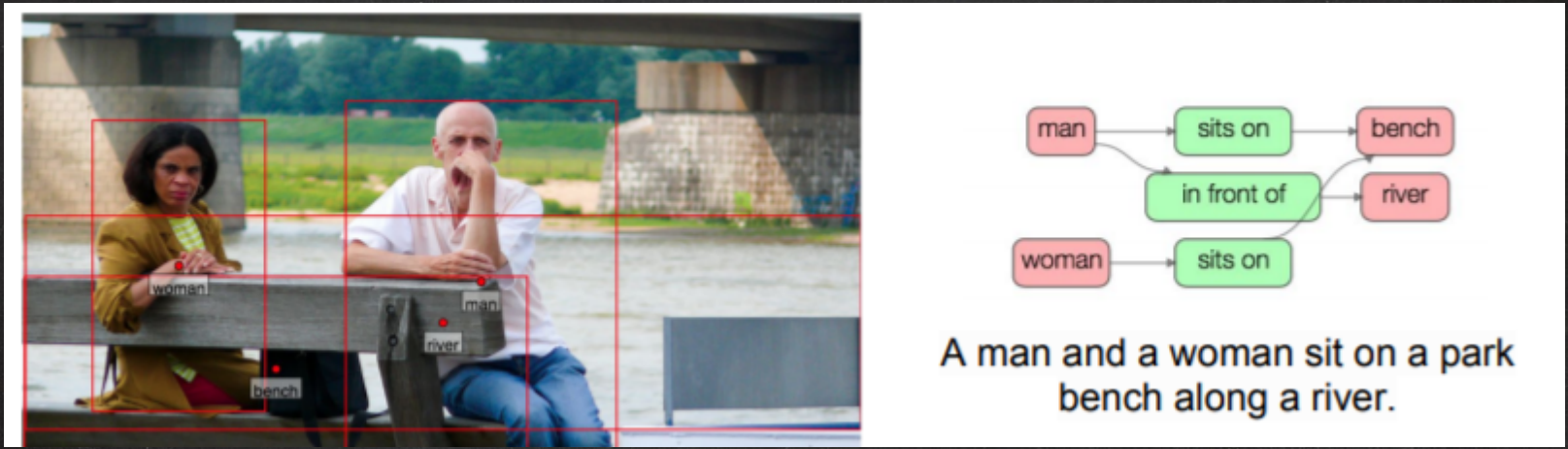
Baseline model: localization without relationship

Full model: localization and relationship

Qualitative results on synthetic dataset

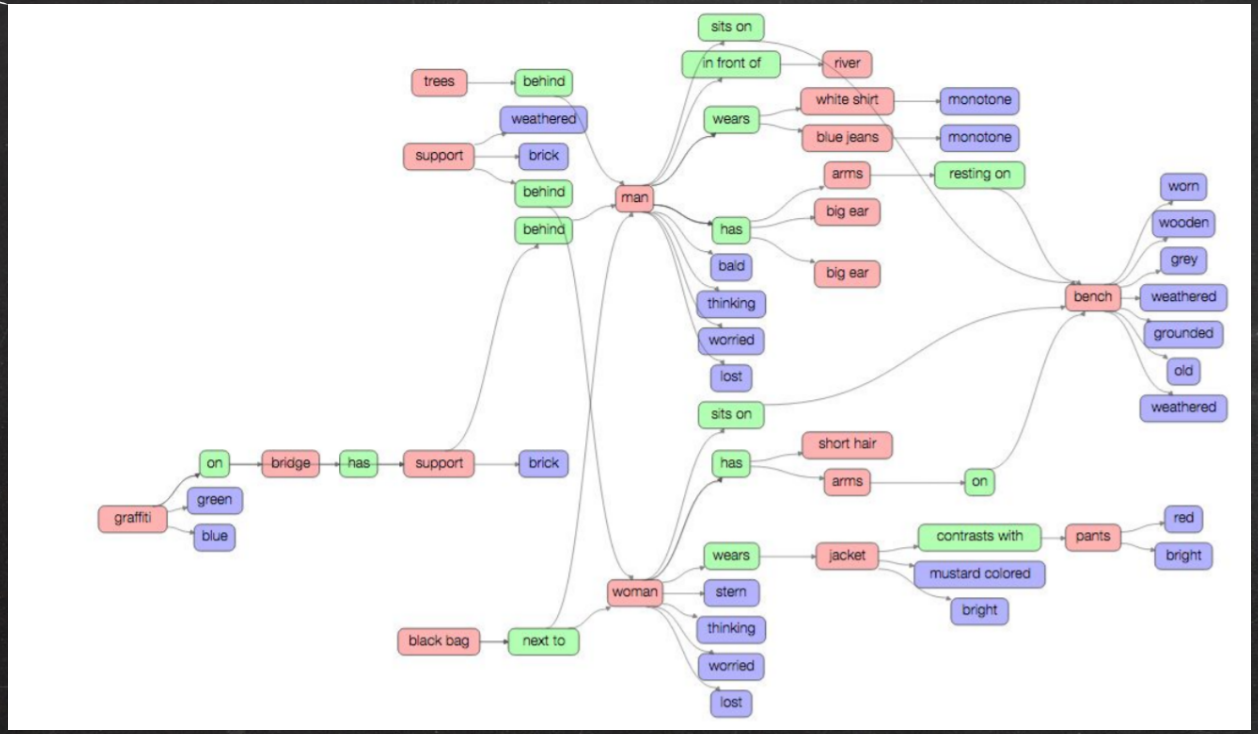


Visual Genome dataset



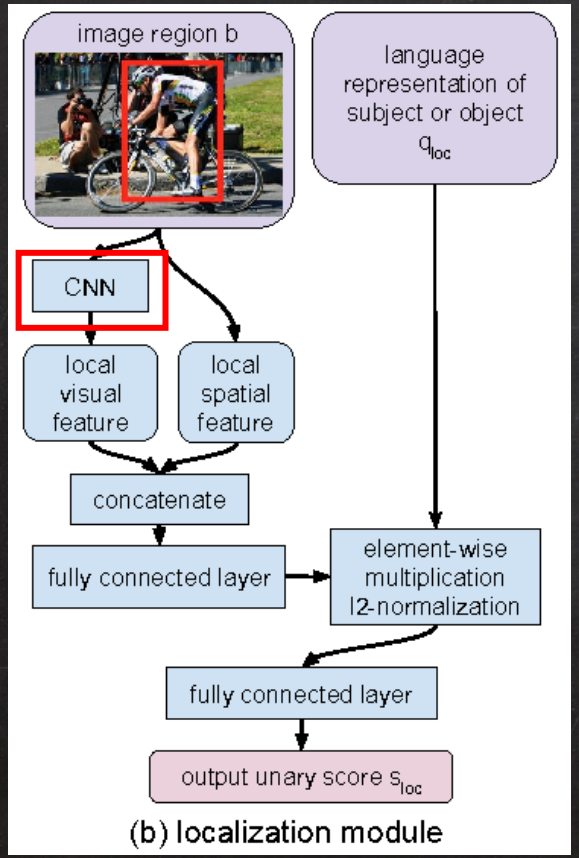
Real images and relationship annotations

Visual Genome dataset



Multiple expressions and bounding boxes on the images

Experiments on Visual Genome dataset



Fc7 output of a Faster-RCNN VGG-16 pretrained on MSCOCO detection dataset

Experiments on Visual Genome dataset

Method	training supervision	P@1-subj	P@1-pair
baseline	subject-GT	41.20%	-
baseline	subject-object-GT	-	23.37%
our full model	subject-GT	43.81%	26.56%
our full model	subject-object-GT	**44.24%**	**28.52%**

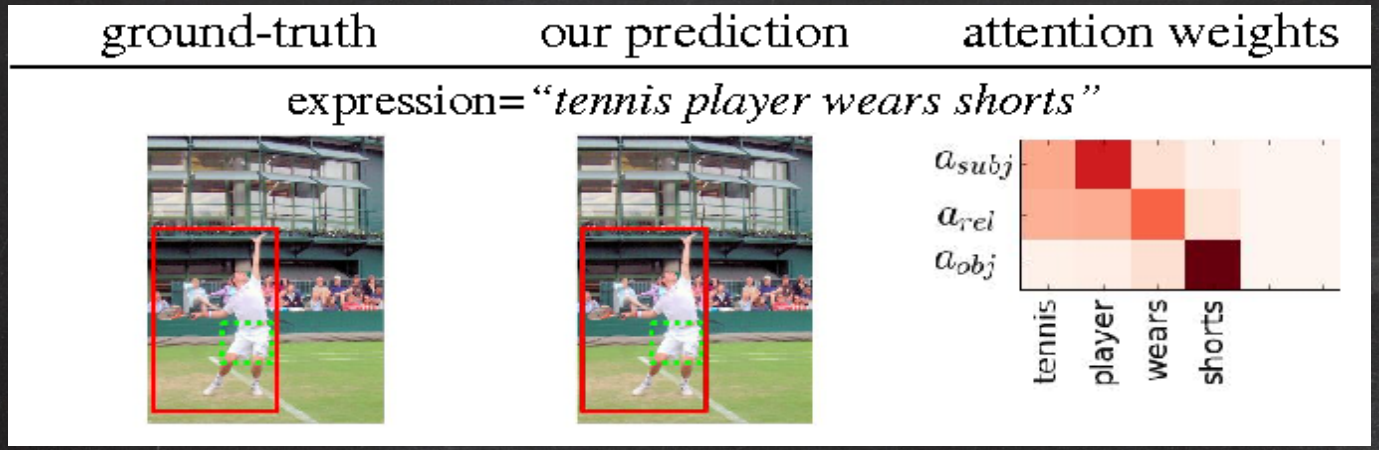
Baseline model: localization without relationship

Full model: localization and relationship

Subject only: find the correct subject

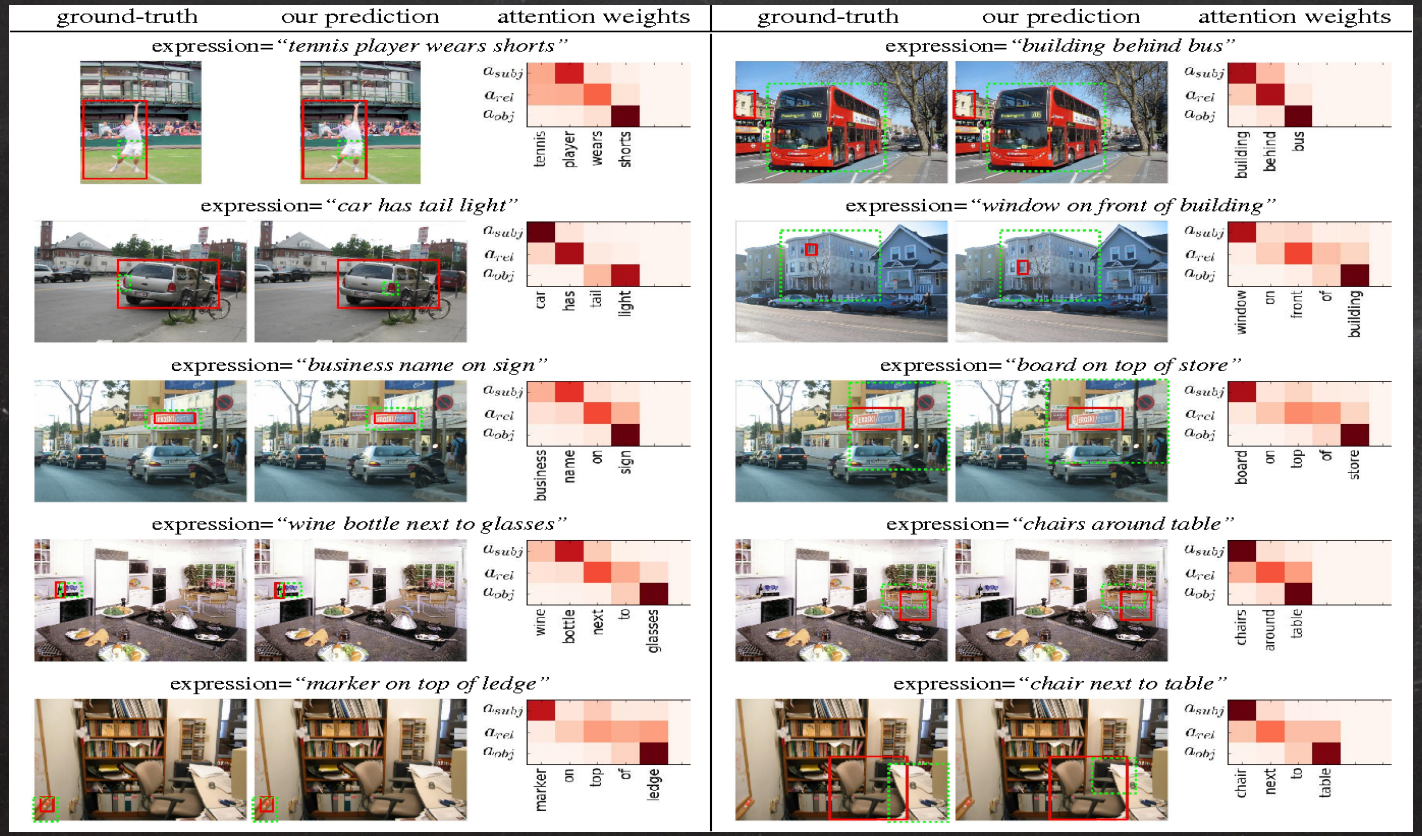
Subject-object: find the correct subject-object pair

Qualitative results on Visual Genome dataset



Performance of language representation

Qualitative results on Visual Genome dataset



Google-Ref dataset

expression="*a bear lying to the right of another bear*"

Real images and relationship expression

Only have ground truth bounding box for the subject

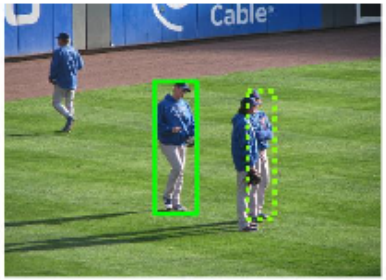
Experiments on Google-Ref dataset

Method	P@1
Mao *et al.* [20]	60.7%
Yu *et al.* [30]	64.0%
Nagaraja *et al.* [21]	68.4%
baseline (loc module)	66.5%
our model (w/ external parser)	53.5%
our full model	**69.3%**

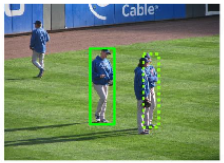
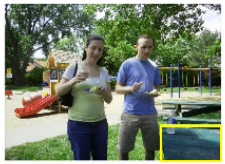
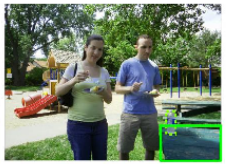
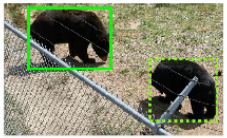
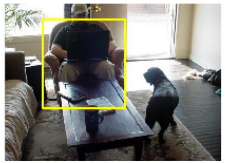
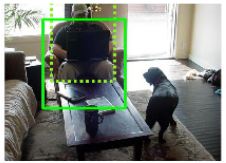
Same configuration on CNN and baseline model

Model (w/ external parser): using external language parser instead of language representation module

Qualitative results on Google-Ref dataset

ground-truth	our prediction	ground-truth	our prediction
expression="*a bear lying to the right of another bear*"		expression="*man in sunglasses walking towards two talking men*"	
			
correct		correct	

Qualitative results on Google-Ref dataset

ground-truth	our prediction	ground-truth	our prediction	ground-truth	our prediction
expression="*a bear lying to the right of another bear*"		expression="*man in sunglasses walking towards two talking men*"		expression="*a picnic table that has a bottle of water sitting on it*"	
					
correct		correct		correct	
expression="*woman in a cream colored wedding dress cutting cake*"		expression="*a man going before a lady carrying a cellphone*"		expression="*pizza slice not eaten*"	
					
correct		correct		incorrect	
expression="*a full grown brown bear near a young bear*"		expression="*black dog standing on all four legs*"		expression="*chair being sat in by a man*"	
					
correct		incorrect		correct	

Visual-7W dataset

Q: What endangered animal is featured on the truck?

A: **A bald eagle.**
A: A sparrow.
A: A humming bird.
A: A raven.

Q: Where will the driver go if turning right?

A: **Onto 24 ¾ Rd.**
A: Onto 25 ¾ Rd.
A: Onto 23 ¾ Rd.
A: Onto Main Street.

Q: When was the picture taken?

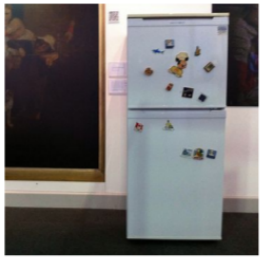
A: **During a wedding.**
A: During a bar mitzvah.
A: During a funeral.
A: During a Sunday church service.

Q: Who is under the umbrella?

A: **Two women.**
A: A child.
A: An old man.
A: A husband and a wife.

Q: Why was the hand of the woman over the left shoulder of the man?

A: **They were together and engaging in affection.**
A: The woman was trying to get the man's attention.
A: The woman was trying to scare the man.
A: The woman was holding on to the man for balance.



Q: How many magnets are on the bottom of the fridge?

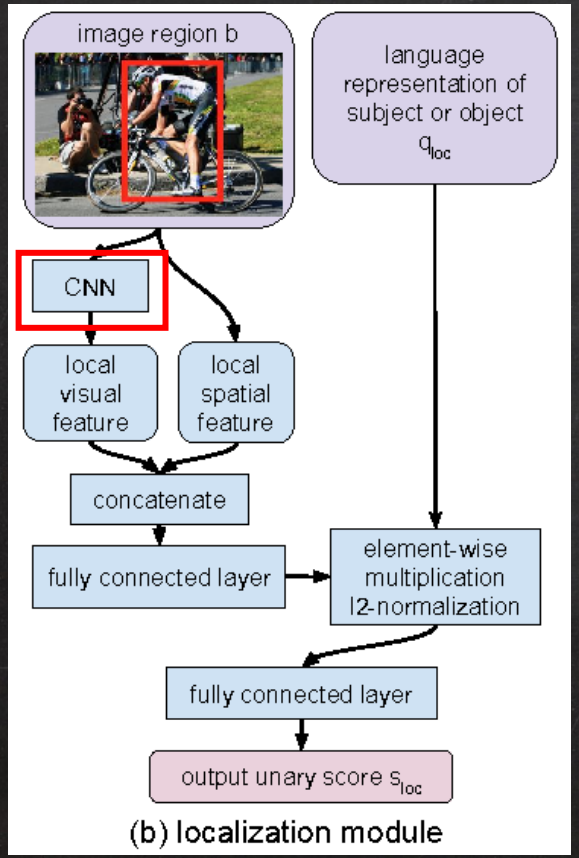
A: **5.**
A: 2.
A: 3.
A: 4.

Visual-7W dataset

question="*Which wine glass is in the man's hand?*"

Multiple choices questions start with which.

Experiments on Visual-7W dataset



Fc7 output of a Faster-RCNN VGG-16 pretrained on MSCOCO detection dataset

RPN in Faster-RCNN to proposal regions for object

Experiments on Visual-7W dataset

Method	Accuracy
Zhu *et al.* [32]	56.10%
baseline (loc module)	71.61%
our model (w/ external parser)	61.66%
our full model	**72.53%**

Same configuration on CNN and baseline model

Model (w/ external parser): using external language parser instead of language representation module

Qualitative results on Visual-7W dataset

question="*Which wine glass is in the man's hand?*"

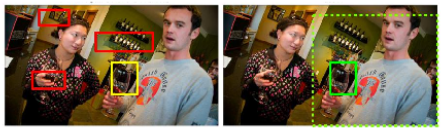
correct

question="*Which person is wearing a helmet?*"

correct

Qualitative results on Visual-7W dataset

question="*Which wine glass is in the man's hand?*"



correct

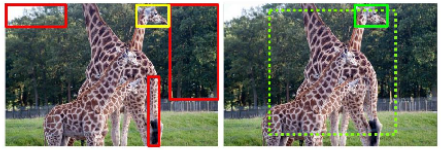
question="*Which person is wearing a helmet?*"

correct

question="*Which mouse is on a pad by computer?*"

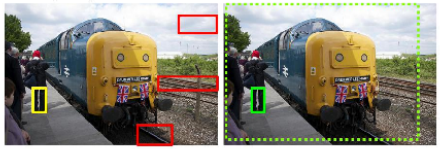
correct

question="*Which head is that of an adult giraffe?*"



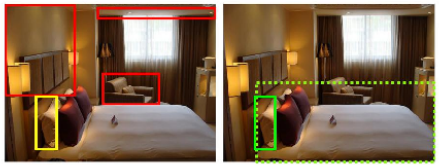
correct

question="*Which pants belong to the man closest to the train?*"



correct

question="*Which white pillow is leftmost on the bed?*"

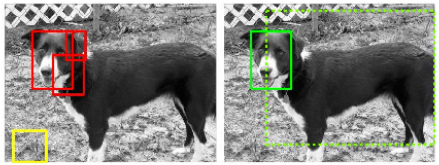


correct

question="*Which red shape is on a large white sign?*"

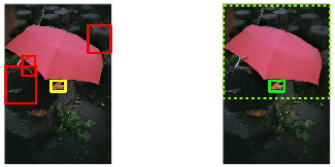
correct

question="*Which is not a pair of a living canine?*"



incorrect

question="*Which hand can be seen from under the umbrella?*"



correct

Comments

Contribution:

- Compositional module for subjects, objects, relations
- Attention mechanism to split the expression

Limitation:

- Need special format of the referential expression

Potential Extensions

Potential Changes:

- Changing relationships to sentences
  (using a dataset with frames and learning just a FC layer for each role)
  (Or learning the frames)

Potential Improvements:

- Better visual feature extractors

Thanks

THANKS!